Page 3, Agent's experience

$$Q(s_t, a_t) = \mathbb{E}[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots + \gamma^k \max_a Q(s_{t+k,a}]$$

To support this situations in a generic way, in PTAN we have class ptan.experience.ExperienceSourceFirs which takes the environment and the agent and provides us the stream of experience tuples: (s_t, a_t, R_t, s_{t+k}), where $R_t = r_t t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots + \gamma^{k-1} r_{t+k-1}$. When k=1, R_t is just the r_t.

Page 10, N-steps DQN

$$Q(s_t, a_t) = r_t + \gamma \max_a Q(s_{t+1}, a_{t+1})$$

$$Q(s_t, a_t) = r_t + \gamma \max_a [r_{a,t+1} + \gamma \max_{a'} Q(s_{t+2}, a')]$$

$$Q(s_t, a_t) = r_t + \gamma r_{t+1} + \gamma^2 \max_{a'} Q(s_{t+2}, a')$$

Page 13, Double DQN

$$Q(s_t, a_t) = r_t + \gamma \max_a Q'(s_{t+1}, a_{t+1})$$

$$Q(s_t, a_t) = r_t + \gamma \max_a Q'(s_{t+1}, \arg\max_a Q(s_{t+1}, a))$$

Page 23, Prioritized replay buffer

From the mathematical point of view, priority of every sample in the buffer is calculated as $P(i) = \frac{p_i^\alpha}{\sum_k p_k^\alpha}$, where p_i is the priority of i-th sample in the buffer and α is the number which shows

To get this, the authors used sample weights, which need to be multiplied to the individual sample loss. The value of weight for each sample is defined as $w_i = (N \cdot P(i))^{-\beta}$, where is another hyperparameter, which should be between 0 and 1.

Page 29, Dueling DQN

This constraint could be enforced in various ways, for example, via the loss function, but in the above paper, the authors proposed a very elegant solution by subtracting from the Q expression in the network mean value of the advantage, which effectively pulls the mean for advantage to the zero: $Q(s, a) = V(s) + A(s, a) - \frac{1}{N} \sum_k A(s, k)$.

Page 32, Categorical DQN

As a next step, the authors have shown that Bellman equation can be generalized for distribution case and it will have a form $Z(x, a) \stackrel{D}{=} R(x, a) + \gamma Z(x', a')$, which is very similar to familiar Bellman equation, but now Z(x, a), R(x, a) are the probability distributions and not numbers. The effect of this equation on distribution is demonstrated on the plots below, taken from the original paper.

We have original distribution of values shown in upper-left corner. After multiplication on γ (upper-right chart)

Page 39

At the end of the function, we need to compute output of network and calculate KL-divergence between projected distribution and networks output for the taken actions. KL divergence shows how much two distributions differ and is defined as $D_{KL}(P\|Q) = -\sum_i p_i \log q_i$